Dynamic Coreference-Based Summarization

Breck Baldwin
Institute for Research
in Cognitive Science
University of Pennsylvania

Thomas S. Morton
Department of Computer
and Information Science
University of Pennsylvania

{breck,tsmorton}@linc.cis.upenn.edu

Introduction

We have developed a query-sensitive text summarization technology well suited for the task of determining whether a document is relevant to a query. Enough of the document is displayed for the user to determine whether the document should be read in its entirety. Evaluations indicate that summaries are classified for relevance nearly as well as full documents. This approach is based on the concept that a good summary will represent each of the topics in the query and is realized by selecting sentences from the document until all the phrases in the query which are represented in the summary are 'covered.' A phrase in the document is considered to cover a phrase in the query if it is coreferent with it. This approach maximizes the space of entities retained in the summary with minimal redundancy. The software is built upon the CAMP NLP system [2].

Problem Statement

Given the relative immaturity of summarization technologies and their evaluation, it is worthwhile to describe our approach in detail and the problems it is intended to solve. An important aspect of our technique is that we produce sentence extraction summaries which are constructed by selecting sentences from the source document. In addition, our summaries are focused on providing relevant information about a query. We feel that the current state-of-the-art techniques are better equipped to produce high quality query-sensitive summaries than generic summaries. Our goal is to produce 'indicative' summaries [4] which allow a user to determine whether the document is relevant to his or her query. The summary is not intended to replace the document or provide answers to questions directly but may have this effect.

Casting our technology in terms of a product, we see the application as an intermediate step between viewing entire documents and the output of an information retrieval engine. Instead of looking at either headlines or an entire document, the user would look at the summaries of the documents and then decide whether the document merited further reading.

Approach

We conducted a simple experiment with summaries produced in the TIPSTER summarization dry run [6]. For 5 queries with 200 documents each, we took the set of summaries produced by the 6 dry-run participants and retained only those summaries that were true-positives, i.e., the summary was judged 'relevant' and the full document was judged 'relevant'. Over all the queries, at least one of the six systems produced a true-positive summary for 96.6% of the documents, although no individual system performed nearly at that level. This meant that some existing technology produced a correct summary for almost every relevant document. Hence we viewed the problem as one of balancing the capabilities of our system to behave like the amalgamated system implicit in joined output. Based on this result we are confident that this class of summarization is tractable with current technologies and this has strongly motivated our design decisions.

Upon encountering a query like "Reporting on possibility of and search for extra-terrestrial life/intelligence.", we assume that the user has defined a class of actions, ideas, and/or entities that he or she is interested in. The job of an information retrieval engine is to find instantiations of those classes in text documents in some database. We view summarization as an additional step in this process where we attempt to present the user with the smallest collection of sentences in the document that instantiate the user specified classes and do not mislead the user about the overall content of the document. By doing so, we can greatly shorten the amount of the document that the user must read in order to determine whether the document is relevant for the user's needs.

Just as information retrieval algorithms approximate document relatedness by examining various string matchings between the query and the text, we approximate certain classes of coreference between the query and the text by examining linguistic information. These coreference relations include identity of reference and part-whole relations for nominal and verbal phrases.[1] This moves us a step closer to reasoning at a more appropriate level of generalization, for summarization, which is still technologically feasible. Below are examples indicating the classes of relatedness that we are trying to capture.

The identity relation between the query and the document

Noun phrase coreference is the best understood class of relations that we compute. For example, there is coreference between 'Federal Emergency Management Agency' in the query and the acronym 'FEMA' in the document below:

> *Query:* What is the main function of the **Federal Emergency Management Agency** and the funding level provided to meet emergencies?
>
> *Document:* ...**FEMA** agrees that "fine-tuning" is needed to the 1974 act establishing a coordinated federal program to prepare for and respond to hurricanes, tornadoes, storms and floods. ...

Since these noun phrases refer to the same entity in the world, sentences that mention the organization would be particularly valuable in a summary. This class of coreference can include people, companies and objects such as automobiles or aluminum siding. It need not be restricted to proper nouns as it is possible to refer to an entity using common nouns, i.e. 'the agency' and pronouns.

Identity also holds between events mentioned in the query and document. Sometimes the event that a query describes is the best indicator of what document should be retrieved, and correspondingly what sentences are appropriate for a summary. Consider the following:

> *Query:* A relevant document will provide new theories about **the 1960's assassination** of President Kennedy.
>
> *Document:* ...The House Assassinations Committee concluded in 1978 that Kennedy was "probably" **assassinated** as the result of a conspiracy involving a second gunman, a finding that broke from the Warren Commission's belief that Lee Harvey Oswald acted alone in Dallas on Nov. 22, 1963. ...

The noun phrase 'the 1960's assassination' refers to an event, which is the same as the one referred to in the document with the verb 'assassinated'. Note also that there is coreference between 'President Kennedy' and 'Kennedy' in the document.

The part-whole relation between the query and the document

In addition to the identity relation, phrases in a text which refer to parts of an entity or concept mentioned in the query will likely provide useful information, and therefore should be included in a summary. Finding these relations in in general is beyond the scope of this paper, however, our approximation of a subclass of these relations proved helpful for a number of queries.

A strong example of the part-whole relation occurs when a country is mentioned in the query and a province or city within that country is mentioned in the document. For example:

> *Query:* Document will discuss efforts by the black majority in **South Africa** to **overthrow** domination by the white minority government.
>
> *Document:* About 90 soldiers have been arrested and face possible death sentences stemming from a coup attempt in **Bophuthatswana**, ...Rebel soldiers **staged** the takeover bid Wednesday, **detaining** homeland President Lucas Mangope. ...

Bophuthatswana is inside South Africa, and sentences that mention it are clearly good candidates for inclusion in a summary.

We also consider part-whole relations between events as in the relation between 'overthrow' and 'staged' and 'detained'. Those events are sub-parts of overthrow events, and as such, sentences that contain sub-parts of the events are reasonable candidates for inclusion in summaries.

Implementation

The summarization technique was developed within the CAMP NLP framework. This system provides an integrated environment in which to access many levels of linguistic information as well as world knowledge. Its main components include: named entity recognition, tokenization, sentence detection, part-of-speech tagging, morphological analysis, parsing, argument detection, and coreference resolution. Many of the techniques used for these tasks perform at or near the

[1] It is not clear whether more sophisticated annotations are appropriate for information retrieval, and perhaps more to the point, it is not clear that there are sufficient resources to process 2 GB collections of data.

state of the art and are described in more depth in [12, 9, 8, 7, 5, 1, 2]. The system produces coreference annotated documents which serve as the input to the summarization algorithm.

Relating the query to the document

The relationships discussed previously are approximated via a series of associations between tokens in the query, headline, and the body of the document. Event references are captured by associating verbs or nominalizations in the query with verbs and nominalizations in the document.

Given three verbal forms v_1 in the query, v_2 in the document, and v_3 in the set of all verbal forms, where a verbal form is the morphological root of a verb or the verb root corresponding to a nominalization, v_1 is associated with v_2 if at least one of the following criteria are met:

1. $(v_1 \neq v_2) \wedge p(v_1, v_2)/(p(v_1)p(v_2)) \geq 5$
2. $(v_1 = v_2) \wedge (\exists v_3 \neq v_1 \mid p(v_1, v_3)/p(v_1)p(v_3) \geq 5)$
3. $(v_1 = v_2) \wedge ((subject(v_1) = subject(v_2)) \vee (object(v_1) = object(v_2)))$

Here $p(v_i)$ is the probability that v_i occurs in a document and $p(v_i, v_j)$ is the probability that v_i and v_j occur in the same document. These probabilities are based on frequencies gathered from approximately 45,000 Wall Street Journal articles. Criterion 1 is a measure of mutual information between two verbs. Criterion 2 is used to rule out frequently occurring verbs such as "be" and "make". Criterion 3 allows for verbs which are ruled out by criterion 2 to be associated when additional context is available. This is important since some queries only contain verbal forms which are ruled out by criterion 2.

Relationships between proper nouns are made on the basis of string matches, acronym matching, and dictionary lookup. Acronyms are determined either through a table lookup or an appositive construction occurring in the document which designates the acronym for a specific proper noun. A proper noun in the query is considered associated with a proper noun in the document if it matches the string or acronym of the proper noun in the document or it appears in the definition of the proper noun in the document. A reverse dictionary lookup often allows cities to be associated with the country they are in.

A token in the query which is a lowercase noun or adjective is associated with any token in the document which matches its morphological root and part of speech.

Tokens which occur in the headline are associated with tokens in the document body using the same criteria as the query, with the exclusion of the dictionary lookup. The dictionary lookup was excluded because the headline will likely use the same lexicalization of a proper noun as that used in a document. This is less likely to be the case with the query.

Selecting a sentence

The associations discussed in the previous section are used to rank and select sentences from the document. Every token in the document which is associated with the same token in the query or headline is considered to be in the same coreference chain. A sentence which contains any token in a given coreference chain is said to cover that chain.

The following scores are computed for each sentence in the document:

1. The number of coreference chains from the query which are covered by the sentence and haven't been covered by a previously selected sentence.
2. The number of noun coreference chains from the query which are covered by the sentence and the number of verbal terms in the sentence which are chained to the query.
3. The number of coreference chains from the headline which are covered by the sentence and haven't been covered by a previously selected sentence.
4. The number of noun coreference chains from the headline which are covered by the sentence and the number of verbal terms in the sentence which are chained to the headline.
5. The number of coreference chains which are covered by the sentence and haven't been covered by a previously selected sentence.
6. The number of noun coreference chains which are covered by the sentence.
7. The index of the sentence in the document; sentences are sequentially numbered.

The sentences are sorted based on the above scores, where the *ith* scoring criteria is only considered in case of a tie for all criteria less than i. Scores 1-6 are ranked in descending order while score 7 is ranked in ascending order. The top-ranked sentence is selected, and scores 1, 3, and 5 are recomputed in order to select the next sentence. Selection halts when all coreference chains in the query have been covered and the summary contains at least 4 sentences.

Scores 1 and 2 are used to select sentences which are related to the query. Scores 3 and 4 are motivated by documents which have 1 or 2 sentences which appear related to the query but if presented alone would give a false impression of the true content of the document. Thus sentences related to the headline are presented to provide additional background. Consider the following example:

> *Query:* What evidence is there of paramilitary activity in the U.S.?
>
> *Summary:* ...Last month the extremists used rocket-propelled grenades for the first time in three attacks on police and paramilitary units. ...

This sentence was selected because it contains tokens which are in coreference chains with tokens in the query; however, alone it is potentially misleading because the place of the attack is not mentioned. This ambiguity is resolved when the following sentence is selected because it is well associated with the headline.

> *Summary:* ...Sikh militants may have acquired one or two U.S.-made Stinger anti-aircraft missiles and hidden them inside the Golden Temple, the Sikh faith's holiest shrine, Punjab police officials said Saturday....

This provides enough background information for the reader to realize that the para-military activity is not taking place in the U.S. and thus that the document is irrelevant to the query.

Likewise, scores 5 and 6 act similarly to 3 and 4 for documents which do not contain a headline. We found this particularly important for advertisements which often don't state a product or company name in the beginning of the document, but will repeat these names numerous times throughout the document.

Generating the summary

Once sentences have been selected, they are presented in the order they occurred in the document. Pronouns which do not have a referent in the previous sentence of the summary are filled with a more descriptive string whenever a referent can be determined. If space is of concern, prepositional phrases attached to nouns (which are not nominalizations), appositives, conjoined noun phrases and relative clauses are removed, provided they contain no tokens associated with the query or the headline. Since determining pronoun referents and the selection of clauses for removal are subject to errors, filled pronouns are placed in square brackets and removed clauses are replaced with an ellipsis to indicate to the reader that the original text has been modified.

Example summary

An example summary which demonstrates many of the features of our system appears below. It has been constrained to be approximately 10% of the original document length, so it is not representative of the summaries used in the evaluation, but it contains examples of the of both pronoun filling and clause deletion.

The last sentence in the summary was selected first because the tokens "death","sentence", "kill", and "term" were associated with the nominalization "punishment". The stranded pronoun "it" has also been filled. Sentence 2 was selected next because of the match-up between the verb "is" and the object "deterrent" in the document and the query. Finally, the first sentence was chosen because there is another mention of the prison name "Marion" in the document. This summary differs from the one generated when the 10% length constraint is not imposed, because some higher ranked sentences were passed over since their inclusion would have exceeded the length restriction.

> *Query:* Is there data available to suggest that capital punishment is a deterrent to crime?
>
> *Summary:* "Marion is basically the end of the line," Bogdan said.
> ... There is no deterrent ... to keep them from doing this again.
> Additionally, [the pending Senate bill] would create five new death penalty offenses: murder by a federal inmate serving a life sentence; drug kingpins in a continuing criminal enterprise even if no murders occur; drug kingpins who try to kill to obstruct justice; drug felons who unintentionally kill with aggravated recklessness; and people who kill with a firearm during a violent ... crime.

Evaluation

In order to evaluate our summarization algorithm, we selected 10 unseen queries from the Text REtrieval Conference (TREC) document collection. Summaries were generated for 200 documents, 20 per query, and assessors[2] were asked to make relevance judgments based on the summaries. A document was considered relevant if it contained the information requested in the query or if the assessor believed that the full document would likely contain this information. The relevance judgments were then compared to those made by the TREC assessors using the full document. This comparison places a summary in one of the following categories:

- a = judged relevant, full document is relevant
- b = judged relevant, full document is irrelevant
- c = judged irrelevant, full document is relevant

[2] Each author served as an assessor making judgments for 100 documents across 10 queries.

- d = judged irrelevant, full document is irrelevant

Precision, recall, and accuracy are then computed as follows:

$$\text{precision} = a/(a+b)$$
$$\text{recall} = a/(a+c)$$
$$\text{accuracy} = (a+d)/(a+b+c+d)$$

Compression is computed over the number of non-whitespace characters in the summary and the original document. Here compression is defined as the percentage of the document that was not included in the summary:

$$\text{compression} = \frac{(length_{document} - length_{summary})}{length_{document}}$$

The results from our experiment are shown in the following table:

Precision	82.8%	101/(101+21)
Recall	77.7%	101/(101+29)
Compression	82.8%	(704686-121272)/704686
Accuracy	75.0%	(101+49)/200

A second evaluation on 910 documents was performed for [4]. These results superficially appear significantly worse than those from the initial evaluation however a more careful analysis (provided in the discussion section) shows that they are in fact similar to the results of the previous evaluation.

Precision	80.3%	322/(322+79)
Recall	57.6%	322/(322+237)
Compression	83.0%	
Accuracy	65.3%	(322+272)/910

Discussion

We view the results of the first evaluation as promising in that they compare favorably with inter-assessor consistency using the entire document. [11] reports unanimous relevance judgments by three assessors for 71.7% of the documents. Interpolating this figure to two assessors yields an 80.1% agreement figure. Using summaries which on average are only 17.2% of the original document, our assessors matched the TREC assessors for 75.0% of the documents.

The second evaluation yielded a much lower recall figure while precision remained comparable. This, however, is also the case when the same assessors judgments on the full documents are compared to those of the TREC assessors. These results are as follows:

Precision	83.5%	167/(167+33)
Recall	63.5%	167/(167+96)
Compression	100.0%	
Accuracy	69.3%	(167+124)/420

We view these results as favorable as well since our accuracy is 65.3% using 17.0% of the document on average compared to 69.3% accuracy using the entire document. The discrepancy between the two evaluations appears to be based on the assessors in the second evaluation using a stricter criteria for relevance than that used by the previous evaluation's assessors or the TREC assessors.

It was noted after the first evaluation that different criteria for relevance accounted for some of the disagreement between our assessors and the TREC assessors. Many documents considered relevant were marked as irrelevant due to different notions of relevance and not because the summary failed to provide material on which to base a correct decision. These difficulties only hinder the evaluation of a summary system and not its use in an application, since a user will have a clear idea of his or her intentions when determining a document's relevance.

As we mentioned previously, our approach has been to balance methods of relating the query to sentences in the document. The nearly 100% recall of the dry-run summaries encouraged us, and we even used the output of those summaries to provide a test-bed for evaluating our summaries. Although we never actively sought to emulate aspects of other systems directly, our final algorithm does share some basic ideas and approaches from those systems. Some of the similarities are listed below:

In [3], they eliminate redundant information from summaries by classifying sentences according to Maximal Marginal Relevance (MMR). MMR ranks text chunks according to their dissimilarity to one another. Summaries can then be produced with sentences that are maximally dissimilar, thereby increasing the likelihood that distinguishing information will be in the summary. One can view our coverage requirement for terms in the query as an attempt to pick dissimilar sentences from the document. Instead of MMR, we use the fact that a sentence which does not contain redundantly referring phrases to the query is more highly ranked than a sentence that does.

Our individual sentence scoring algorithm shares some properties with [10]. Their approach includes scores for anaphoric density, string equivalence with the title or headline of a document, and position of the sentence in the document. However, we do not take advantage of overt cues for summary sentences, such as 'in summary' or 'in conclusion', nor do we use temporal information in generating a summary.

Like many systems, we do a form of word expansion in attempting to relate the query to the document. However, the fact that we restrict expansion to proper nouns and verbs and their nominalizations is notable. We found this limited set of expansions restricts the relations between the text and the query well and also fits within the framework of part-whole relations in coreference. We did not consider part-whole relations for common nouns, because in practice we have not had very good results limiting over-generation in that domain.

Conclusions and Future Work

We have developed and tested a query-sensitive text summarization system that is nearly as effective as full text documents for determining whether a document is relevant to the query. The system uses a limited class of coreference-based relations between the query and the document to select sentences which represent instantiations of entities, events, or concepts articulated in the query. The algorithm is implemented within the CAMP NLP system and utilizes linguistic generalizations like part-of-speech, parsing and predicate-argument structure.

An issue in evaluating our system is that the input data has been selected by an information retrieval engine. As such, we have no data on how well our summaries would work on relevant documents that the information retrieval engine fails to retrieve. These engines tend to select documents based on string matching and we have shown that our summarization technology does an excellent job of summarizing them. However, the information retrieval engine may be acting as an advantageous filter on the space of documents. It would be interesting to do experiments on relevant documents that contain very few string matches with the query.

In the future we hope to improve the accuracy of the coreference relations. Specifically, we will focus on the recognition of events which we believe are very important to a large class of queries.

Acknowledgments

We would like to acknowledge three anonymous reviewers for their helpful comments and Tonia Bleam for providing assessments during the development of this system.

References

[1] Breck Baldwin. CogNIAC: High precision coreference with limited knowledge and linguistic resources. In *Proceedings of the ACL Workshop on Operational Factors in Practical, Robust Anaphora resolution for Unrestricted Texts*, pages 38–45, Madrid, Spain, June 1997.

[2] Breck Baldwin, Christine Doran, Jeffrey C. Reynar, Michael Niv, B. Srinivas, and Mark Wasson. EAGLE: An extensible architecture for general linguistic engineering. In *Proceedings of RIAO-97*, Montreal, 1997.

[3] Michael Bett and Jade Goldstein. Automated query-relevant document summarization. In *Proceedings of Tipster Text Phase III 12-Month Workshop*, 1997.

[4] Michael Chrzanowski, Therese Firmin, Lynette Hirschman, David House, Inderjeet Mani, Leo Obrst, Sara Shelton, Beth Sundheim, and Sandra Wagner. (SUMMAC) call for participation. http://www.tipster.org/summcall.htm, January 1998.

[5] Michael John Collins. A New Statistical Parser Based on Bigram Lexical Dependencies. In *Proceedings of the 34th Annual Meeting of the ACL*, 1996.

[6] Therese Hand. Tipster summarization evaluation task:dry-run evaluation results. In *Proceedings of Tipster Text Phase III 12-Month Workshop*, 1997.

[7] Daniel Karp, Yves Schabes, Martin Zaidel, and Dania Egedi. A freely available wide coverage morphological analyzer for english. In *Proceedings of the 15th International Conference on Computational Linguistics*, 1994.

[8] Adwait Ratnaparkhi. A Maximum Entropy Part of Speech Tagger. In Eric Brill and Kenneth Church, editors, *Conference on Empirical Methods in Natural Language Processing*, University of Pennsylvania, May 17-18 1996.

[9] Jeffrey C. Reynar and Adwait Ratnaparkhi. A maximum entropy approach to identifying sentence boundaries. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 16–19, Washington, D.C., April 1997.

[10] Tomek Strzalkowski, Fang Lin, Jin Wang, Langdon White, and Bowden Wise. Natural language information retrieval and summarization. In *Proceedings of Tipster Text Phase III 12-Month Workshop*, 1997.

[11] Ellen M. Voorhees and Donna Harman. Overview of the fifth Text REtrieval Conference (TREC-5). In *Proceedings of the Fifth Text REtrieval Conference (TREC-5)*, pages 1–28. NIST 500-238, 1997.

[12] Nina Wacholder, Yael Ravin, and Misook Choi. Disambiguation of proper names in text. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, May 1997.